2011 International Conference on Information and Network Technology
IPCSIT vol.4 (2011) © (2011)IACSIT Press, Singapore

Automatic Evaluation of Cluster in Unlabeled Datasets

M.Krishnamoorthi [1]
[1]Assistant Professor, Dept. of CSE.,
Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India.
(krishna_bit@yahoo.co.in)

Abstract - All clustering algorithms ultimately rely on one or more human inputs, and the most important input is number of clusters (c) to seek. There are "adaptive" methods which claim to relieve the user from making this most important choice, but these methods ultimately make the choice by thresholding some value in the code. Thus, the choice of 'c' is transferred to the equivalent choice of the hidden threshold that determines 'c' automatically. This work investigates a new technique called Spectral VAT for estimating the number of clusters to look for in unlabeled data utilizing the VAT [Visual Assessment of Cluster Tendency] algorithm, coupled with a Spectral analysis and several common image processing techniques. Several numerical datasets are presented to illustrate the effectiveness of the proposed method.

Keywords - Cluster tendency, Reordered dissimilarity image, VAT, SpecVAT.

1. Introduction

Cluster analysis is the problem of partitioning a set of objects O = $\{O_1 \ldots O_n\}$ into 'c' self-similar subsets based on available data. In general, clustering of unlabeled data posses three major problems. They are 1) assessing cluster tendency, i.e., how many clusters to seek 2) Partitioning the data into 'c' meaningful groups and 3) validating the c clusters that are discovered. We address the first problem, i.e., determining the number of clusters 'c' prior to clustering. Many clustering algorithms require number of clusters as an input parameter, so the quality of the resulting clusters mainly depends on this value. Most methods are post clustering measures of cluster validity i.e., they attempt to choose the best partition from a set of alternative partitions.

In contrast, tendency assessment attempts to estimate 'c' before clustering occurs. Our focus is on preclustering tendency assessment. Here, we represent the structure of the unlabeled data sets as a Reordered Dissimilarity Image (RDI), where pair wise dissimilarity information about a data set including n objects is represented as nxn image. RDI is generated using VAT and highlights potential clusters as a set of dark blocks along the diagonal of the image. So, number of clusters can be easily estimated using the number of dark blocks across the diagonal. The existing technique for preclustering assessment of cluster tendency is Dark Block Extraction (DBE) which obtains less accurate and less reliable results. It does not concentrate on the perplexing and overlaping issues.

The proposed Spectral VAT method combines several common image and signal processing techniques. First, a weighted matrix is calculated the normalised Laplacian matrix is formed and finally RDI that portrays a potential cluster structure from the pairwise dissimilarity matrix of the data is created. For concreteness, we will generate RDIs using the Visual Assessment of Cluster Tendency (VAT) algorithm. Then, sequential image processing operations (region segmentation, directional morphological filtering and distance transformation) are used to segment the regions of interest in the RDI and to convert the filtered image into a distance-transformed image. Finally, we project the transformed image onto the diagonal axis of the RDI, which yields a one-dimensional signal, from which we can extract the (potential) number of clusters in the data set using sequential signal processing operations like average smoothing and peak detection. The proposed method is easy to understand and implement, and encouraging results are achieved on a variety of artificially generated and real-world data sets.

The remainder of this paper is organized as follows: In Section 2, we review related work. Section 3 describes the DBE approach. Section 4 contains a description about the Spectral analysis. Here, we compare SpecVAT to a predecessor algorithm called Dark Block extraction algorithm pointing out similarities and differences between the two approaches.

2. Related work

The selection of the number of clusters is an important and challenging issue in cluster analysis. A number of attempts have been made to estimate ‘c’ in a given data set. Most methods are postclustering measures of cluster validity, i.e., they attempt to choose the best partition from a set of alternative partitions. In contrast, tendency assessment attempts to estimate ‘c’ before clustering occurs. Our focus is on preclustering tendency assessment, but for completeness, we briefly summarize some existing approaches to the postclustering cluster validity problem, before describing visual methods for cluster tendency assessment. Index-based methods for cluster validity usually emphasize the intracluster compactness and intercluster separation and consider the effects of other factors such as the geometric or statistical properties of the data. Ling, (1973) proposed SHADE [16] approach. SHADE approximates what is now regarded as a nice digital image representation of clusters using a crude 15 level halftone scheme created by overstriking standard printed characters. SHADE displays the lower triangular part of the complete square display. Visual identification of (triangular) pattems is considerably more difficult than when a full, square display is used. SHADE was used after application of a hierarchical clustering scheme, as an alternative to visual displays of hierarchically nested clusters via the standard dendrogram.

Probabilistic indices of cluster validity attempt to validate the number of clusters found by probabilistic clustering algorithms. Dellaert, (2002) proposed the ubiquitous Expectation-Maximization (EM) algorithm which produces probabilistic partitions of object data at requested values of ‘c’ . This method amounts to fitting the data with a probabilistic model (e.g., a Gaussian mixture model). Guo et al, (2002) proposed a cluster number selection method for a small set of samples using a Bayesian Ying-Yang (BYY) model. A key limitation of statistical approaches is the need to make (often unrealistic) distributional assumptions about the data. Under the second-order approximation, a new equation for estimating the smoothing parameter in the cost function is derived. Finally, a gradient descent smoothing parameter estimation approach is proposed that avoids complicated integration procedure and gives the same optimal result.

Bezdek [11] proposed a Visual method for cluster tendency assessment (VAT) algorithm. Here, the objects are reordered and the reordered matrix of pair wise object dissimilarities is displayed as an intensity image. Clusters are indicated by dark blocks of pixels along the diagonal. The Visual Assessment of (cluster) Tendency (VAT) method readily displays cluster tendency for small data sets as grayscale images, but is too computationally costly for larger data sets. Bezdek [6] revised version of VAT (reVAT) is presented here that can efficiently be applied to larger collections of data. The bigVAT [9] and sVAT [8] offered different ways to approximate the VAT RDI for very large data sets. The coVAT extended the idea of RDIs to rectangular dissimilarity data to enable tendency assessment for each of the four coclustering problems associated with such data.

Sanghamitra [4] proposed a new symmetry based genetic clustering algorithm is proposed which automatically evolves the number of clusters as well as the proper partitioning from a data set. Strings comprise both real numbers and don’t care symbol in order to encode a variable number of clusters. Here, assignment of points to different clusters is done based on point symmetry based distance rather than the Euclidean distance. A newly proposed point symmetry based cluster validity index, Sym-index, is used as a measure of the validity of the corresponding partitioning. The algorithm is therefore able to detect both convex and non-convex clusters irrespective of their sizes and shapes as long as they possess the symmetry property.

In contrast to the previous version, major changes of this paper are summarized as follows:

- We modify the organization of the paper for better readability, as well as describing each of the proposed algorithms in more detail, including the theoretical analysis of runtime complexity.
- We provide several additional real-world data sets to evaluate our previously proposed algorithms.
- We propose a new strategy to extend the previous algorithms to make them feasible for use with truly large data sets.
- We provide extensive experiments on several synthetic and real data sets to evaluate this new algorithm, and the results demonstrate its effectiveness.

In sharp contrast to index-based validation of postclustering results, in the SpecVAT method we provide an estimate for ‘c’ prior to clustering.

3. Dark Block Extraction

The technique that is used presently for automatically determining the number of clusters in unlabeled data sets is the Dark Block extraction, which is nearly a parameter free method. In short, DBE is an algorithm that counts the dark blocks along the diagonal of a Reordered Dissimilarity Image.

3.1 Steps in DBE Algorithm:

Step 1: Find the threshold value α and D using Otsu's algorithm.

Step 2: Transform D into new dissimilarity matrix **D'** with $\mathbf{D'_{ij} = 1- \exp (-d_{ij}/\alpha)}$

Step 3: Form an RDI image $\mathbf{I}^{(1)}$ using the previous module.

Step 4. Threshold $\mathbf{I}^{(1)}$ to obtain a binary image $\mathbf{I}^{(2)}$ using algorithm of Otsu.

Step 5. Filter $\mathbf{I}^{(2)}$ using morphological operations to obtain a filtered binary image $\mathbf{I}^{(3)}$

Step 6. Perform a distance transform on $\mathbf{I}^{(3)}$ to obtain a gray scale image $\mathbf{I}^{(4)}$ and scale the pixel values to [0, 1].

Step 7. Project the pixel values of the image on to the main diagonal axis of $\mathbf{I}^{(4)}$ to form a projection signal $\mathbf{H}^{(1)}$

Step 8. Smooth the signal $\mathbf{H}^{(1)}$ to obtain the filtered signal $\mathbf{H}^{(2)}$ by an average filter.

Step 9. Compute the first order derivative of $\mathbf{H}^{(2)}$ to obtain $\mathbf{H}^{(3)}$

Step10. Find peak position $\mathbf{p_i}$ and valley positions $\mathbf{v_j}$ in $\mathbf{H}^{(3)}$

Step 11.Select major peaks and valleys by removing the minor ones.

4. SpecVAT

Our work is built upon the VAT algorithm. Two important points about VAT are noted here: 1) Only a pairwise dissimilarity matrix D is required as the input. When vectorial forms of objects are available, it is easy to convert them into D using some form of dissimilarity measures. Even when vectorial data are not explicitly available, it is still feasible to use certain flexible metrics to compute a pairwise dissimilarity matrix, e.g., using Dynamic Time Warping (DTW) to match sequences of different lengths. 2) Although the VAT image suggests both the number of and approximate members of object clusters, matrix reordering produces neither a partition nor a hierarchy of clusters. It merely reorders the data to reveal its hidden structure, which can be viewed as illustrative data visualization for estimating the number of clusters prior to clustering. However, hierarchical structure could be detected from the reordered matrix if the diagonal sub blocks exist within larger diagonal blocks.

At first glance, a viewer can estimate the number of clusters 'c' from a VAT image by counting the number of dark blocks along the diagonal if these dark blocks possess visual clarity. However, this is not always possible. We find that a dark block appears in the VAT image only when a tight (or ellipsoidal) group exists in the data. For complex shaped data sets where the boundaries between clusters become less distinct due to either significant overlap or irregular geometries between different clusters, the resulting VAT images will degrade. Accordingly, viewers may deduce different numbers of clusters from such poor-quality images or even cannot estimate 'c' at all. This naturally raises a problem of whether we can transform D into a new form D' so that the VAT image of D' can become clearer and more informative about the cluster structure. In this paper, we address this problem by combining the VAT algorithm with spectral analysis of the proximity matrix of the data. These spectral methods generally use the eigenvectors of a graph's adjacency (or Laplacian matrix) to construct a geometric representation of the graph. Laplacian Eigenmaps are very similar to the mapping procedure used in a spectral clustering algorithm.

4.1 Steps in SpecVAT

1. Compute the local scaling parameter σ_i for object O_i
2. Construct the weighting matrix W
3. Construct the normalized Laplacian matrix L
4. Choose the k largest eigenvectors of L' to form the matrix V
5. Normalize the rows of V with unit Euclidean norm to generate V'

6. Construct a new pairwise dissimilarity matrix D'
7. Apply the VAT algorithm to D'

The spectral decomposition of the Laplacian matrix provides useful information about the properties of the graph. It has been shown experimentally that natural groups in the original data space may not correspond to convex regions, but once they are mapped to a spectral space spanned by the eigenvectors of the Laplacian matrix, they are more likely to be transformed into tight clusters. Based on this observation, we wish to embed D in a k-dimensional spectral space, where k is the number of eigenvectors used, such that each original data point is implicitly replaced with a new vector instance in this new space. After a comprehensive study of recent spectral methods, we adopt a combination of adjacency graph, weighting function, and graph Laplacian for obtaining a better graph embedding.

5. Discussions & Conclusion

DBE provides an initial estimation of the cluster number, thus avoiding the requirement of repeatedly running a clustering algorithm multiple times over a wide range of ‘c’ in an attempt to find useful clusters.

Table 1 . Results using DBE and SpecVAT

Datasets	**Wine**	**Iris**
# Instances	178	150
# Attributes	13	4
DBE	5	3
SpecVAT	7	4

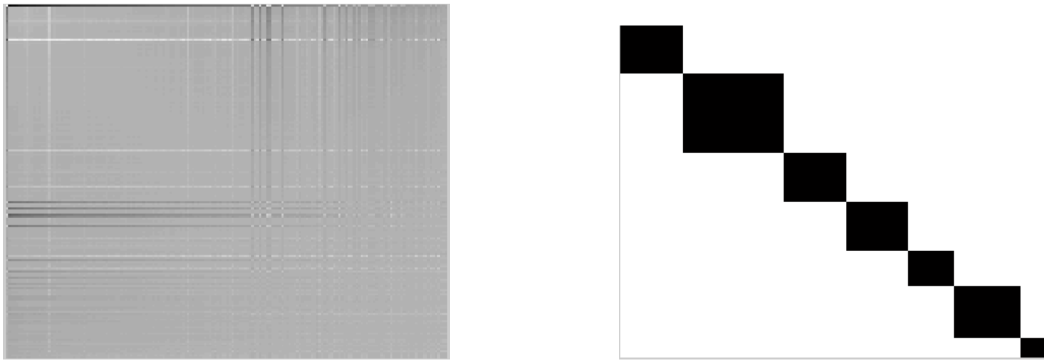

Fig.1 (a) VAT Image of Wine Dataset

(b) Total number of clusters formed for Wine dataset

The resulting Spectral VAT (SpecVAT) images can clearly show the number of clusters 'c' and the approximate sizes of each cluster for data sets with highly irregular cluster structures. Based on spectral VAT, the cluster structure in the data is reliably estimated by visual inspection. In order to better reveal the hidden cluster structure, especially for complex-shaped data sets, the VAT algorithm has been improved by using spectral analysis of the proximity matrix of the data. Based on spectral VAT, a "goodness" measure of SpecVAT images has been proposed for automatically determining the number of clusters. Also, we can derive a visual clustering algorithm based on SpecVAT images and its unique block-structured property.

6. References

[1] L.Waing , Geng , J.Bezdek and C.Leckie , "Enhanced Visual Analysis for Cluster tendency assessment and Data Partitioning" , IEEE Transaction on Knowledge and Data engineering , vol.22, no.10, pp.1401-1414, 2010.

[2] L.Waing, C.Leckie and J.C.Bezdek, "Automatically Determining the Number of Clusters in Unlabeled Datasets", IEEE Transaction on Knowledge and Data engineering, vol. 21, no. 3, 2009.

[3] T.Havens, J.C.Bezdek, J.Keller and M.Popesu, "Dunn's Cluster Valid index as a Contrast Measure of VAT Images", IEEE, 2008.

[4] Sanghamitra Bandyopadhyay and Sripama saha, "A Point Symmetry based Clustering Technique for Automatic Evolution of clusters", IEEE Transaction on Knowledge and Data engineering, vol. 20, no. 11, 2008.

[5] I.Sledge, J.Huband and J.C.Bezdek, "(Automatic) Cluster Count Extraction from Unlabeled Datasets", Joint Proceedings Fourth Int'l Conference Natural Computation (ICNC) and Fifth Int'l Conference on Fuzzy Systems and Knowledge discovery (FSKD), 2008.

[6] J.C.Bezdek, R.J.Hathway and J.Huband, "Visual Assessment of Fuzzy Clustering Tendency for Rectangular Dissimilarity Matrices", IEEE Transactions on Systems, vol. 15, no. 5, pp. 890-903, 2007.

[7] L.Waing and Y. Zhang , "On fuzzy cluster validity indices", Fuzzy Sets and Systems, vol. 158, no. 19, pp. 2095–2117, 2007.

[8] R.Hathway, J.C.Bezdek and J.Huband, "Scalable Visual Assessment of Cluster Tendency", Pattern Recognition, vol.39, no. 6, pp. 1315-1324, 2006.

[9] J.Huband, J.C.Bezdek and R.Hathway, "BigVAT: Visual Assessment of Cluster Tendency ", Pattern Recognition, pp. 1875-1886, 2005.

[10] Gautam Garai, B.B.Chaudhuri, "A Novel Genetic Algorithm for Automatic Clustering", Pattern Recognition, Science Direct Letters 25, pp. 173–187, 2004.

[11] J.C.Bezdek and R.Hathway, "VAT: A tool for Visual Assessment of (Cluster) Tendency", Proceedings of IEEE Int'l Joint Conference on Neural Networks (IJCNN'02), vol. 21, pp. 2225-2230, 2002.

[12] U.Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1650-1654, 2002.

[13] J.C.Bezdek and N.R.Pal, "Some new indexes of cluster validity", IEEE Transactions on Systems, Man, And Cybernetics, vol. 28, pp. 301–315, 1998

[14] J.G.Milligan and M.Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set", Psychometrika, vol. 50, pp. 159-179, 1985.

[15] N.Otsu, "A Threshold Selection Method from Gray-level Histograms", IEEE Transaction on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66,1979.

[16] R.F. Ling, "A Computer Generated Aid for Cluster Analysis,"Comm. ACM, vol. 16, pp. 355-361, 1973.

[17] P. Sneath, "A Computer Approach to Numerical Taxonomy", J. General Microbiology, vol. 17, pp. 201-226, 1957.

[18] R.B. Cattell, "A Note on Correlation Clusters and Cluster Search Methods," Psychometrika, vol. 9, no. 3, pp. 169-184, 1944.